Supplementary Material

Part 1: How to use the RDI (Related Disease Identifier)

The Related Disease Identifier is a knowledge source that contains diseases, syndromes, and symptoms that are known to cause the abnormal laboratory signals (ALS) in our study. This knowledge source was developed by two physicians, in order to facilitate computer executable screening of records. We are providing a sample to facilitate better understanding of our methods. Potential collaborators are encouraged to contact our group to request the full list, along with requirements for appropriate use.

As a standardized vocabulary, we used the freely available National Library of Medicine (NLM) Unified Medical Language System (UMLS) Metathesaurus (license required). From http://www.nlm.nih.gov/pubs/factsheets/umls.html (accessed 11/21/11):

"The Metathesaurus is a very large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health-related concepts, their various names, and the relationships among them. It is built from the electronic versions of many different thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing and cataloging biomedical literature, and/or basic, clinical, and health services research."

The UMLS concepts (CUI values) of interest are grouped according to their associated laboratory abnormality. Below we have provided ten CUIs and the preferred terms for each ALS.

Agranulocytosis:

C0023418	Leukemia
C0852709	Leukemic Phase of Non-Hodgkin's Lyphoma
C0746882	Chronic neutropenia
C0010823	Cytomegalovirus Infection
C0011311	Dengue Fever
C0276275	Disease due to Parvoviridae
C0024141	Lupus Erythmatosus, Systemic
C0001175	Acquired Immunodeficiency Syndrome
C0019159	Hepatitis A
C0018133	Graft-vs-Host Disease

Elevated Creatine Kinase:

C0006434	Burn injury
C0009951	Convulsions
C0036572	Seizures
C0011633	Dermatomyositis
C0027051	Myocardial Infarction
C0155626	Acute myocardial infarction
C1536220	STEMI
C1536221	Non ST segment elevation myocardial infarction
C0013146	Drug abuse
C0085639	Falls

Once the CUIs are chosen, they are then used to isolate the patients who do not have a disease cause for the ALS. All of our free text patient discharge summaries are parsed by our NLP system, MedLEE. For each patient, we have a list of his or her current diseases and symptoms in a standardized form (CUIs) stored in a database table called CODES04_09. We also have table that contains all of the patients with an abnormal laboratory signal, named LABS04_09. To exclude the patient records where a patient has a disease in our RDI list, we use the following example code:

```
select distinct b.mrn, b.rdate, b.code from
(select distinct a.mrn, a.rdate, a.code from LABS04_09 c, CODES04_09 a
where c.CK = 'A' and a.MRN = c.MRN and a.RDATE = c.RDATE) as b,
(select distinct m.mrn, m.rdate, m.code from LABS04_09 d, CODES04_09 m
where d.CK = 'A' and d.MRN = m.MRN and d.RDATE = m.RDATE) as t
where b.MRN = t.MRN and b.RDATE = t.RDATE and
b.code not in
('C0006434',
'C0009951',
```

'C0036572',
'C0006434',
'C0011633',
'C0027051',
'C0155626',
'C1536220',
'C1536221',
'C0013146',
'C0085639',
)
with ur;

From this we have a list of patient medical record numbers that we use to pull the corresponding discharge summary for that unexplained abnormal laboratory signal. These patients are manually reviewed to determine the cause of the ALS.

Part 2: How to develop your own RDI

The RDI is in essence an executable knowledge source. It relies on a knowledge engineer/domain expert to appropriately identify the diseases, syndromes, and symptoms that would result in the Adverse Event. It parallels the physician's differential diagnosis, a systematic diagnostic method, used in patient care. *(There are many ways to grow such a knowledge source, and in so doing, we recommend that others chose the method that recognizes and respects their own domain expert's preferences.)* We would like to share our preferred method for your consideration.

Part A: Physician Recall of Concepts

We start with a collaborative effort, whereby two physicians recall diseases that can cause the laboratory abnormality, in essence compiling a differential diagnosis. Depending on their domain knowledge or area of specialization, they may consult with biomedical literature. This is the similar to the cognitive process that would be used for chart review to identify ADRs, where one asks, "Do I think that this lab abnormality is a reaction to a drug or is there another reasonable explanation for this finding". Instead we are asking, "If we saw a patient with this laboratory abnormality, and this disease, would we conclude that the disease was responsible?"

The generated list can then be coded into standardized UMLS terms or used as a cognitive preparation for the next step. Our personal experience is that stopping at the recall step produces inferior results. *[Unpublished findings: 27 concepts were identified using this method for elevated CK. Testing the use of these concepts as the RDI resulted in a sensitivity of 60% and a specificity of 81%, when the automated patient classification was compared to manual classification for 100 patients.]*

Part B: Physician Recognition of Concepts

We use a script to generate a list of all of the disease and symptom concepts that are related to the adverse event. This list is created in order to facilitate recognition of different degrees of concept granularity by our domain experts. A consequence of the richness of human language contained in free text electronic health records, is that one finds multiple terminology variants for presenting the same concept. In natural language processing, documented diseases and symptoms are automatically coded to corresponding Unified Medical Language System concept unique identifiers.

A domain expert may want to identify all the patients that are hospitalized for any type of myocardial infarction (a concept). This concept may only be documented in the patient chart as a specific type of myocardial infarction, for example "NSTEMI" (non-ST elevation myocardial infarction). The same way a physician reviewing a patient chart is able to recognize that this patient has a concept of interest; these concepts can also be selected from an itemized list.

The list output from our script contains every disease or symptom code present in our adverse event of interest patient group. We calculate the odds ratio for the disease and symptom concepts in the patients with the ALS compared to patients without the ALS and output a ranked list, regardless of statistical significance level. Viewing the list in such a manner is simply our preferred method, but we do not have any empiric evidence that this is more valuable than any other way to do it.

We found 4193 distinct terms (CUIs) associated with elevated creatine kinase.
We found 2916 distinct terms (CUIs) associated with agranulocytosis.

It is challenging to identify an average reading speed for physicians. But in an attempt to put this in context, a study of medical students found that they read at a rate of 100-150 words per minute.[1] At the lower end of this range, one could conclude that the average medical student could read our list in 71 minutes.

There were 175 terms (CUIs) selected for the elevated creatine kinase RDI list
There were 186 terms (CUIs) selected for the agranulocytosis RDI list

References:

1. Klatt EC, Klatt CA. How much is too much reading for medical students? Assigned reading and reading rates at one medical school. Acad Med. 2011 Sep;86(9):1079-83.